

About corpus linguistics, variation, and the variationist method

Benedikt Szmrecsanyi

KU Leuven

Quantitative Lexicology and Variational Linguistics

New Ways of Analyzing Variation 44, Toronto, October 2015



slides @ <http://www.benszm.net/NWAV.pdf>



Introduction



Summary

- because **language variation & change (LVC) work** draws on collections of naturalistic speech, LVC analysts use the corpus-linguistic method
- conversely, many corpus analysts use the variationist method and engage in **corpus-based variationist linguistics (CVL)**
- aim: discuss styles and practices setting apart CVL from LVC; highlight cross-pollination potential

1. LVC in the big picture
2. Corpus-based variationist linguistics (CVL) versus LVC
3. Cross-pollination potential

LVC in the big picture

Corpora and corpus linguistics

“a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis”

(Kennedy 1991: 1)

Intersections and set theory

“a corpus view of language is a view in which texts upon which linguistic analysis is conducted are seen as parts of a larger set of texts or parts of a larger set of texts which can be analysed in a similar way”

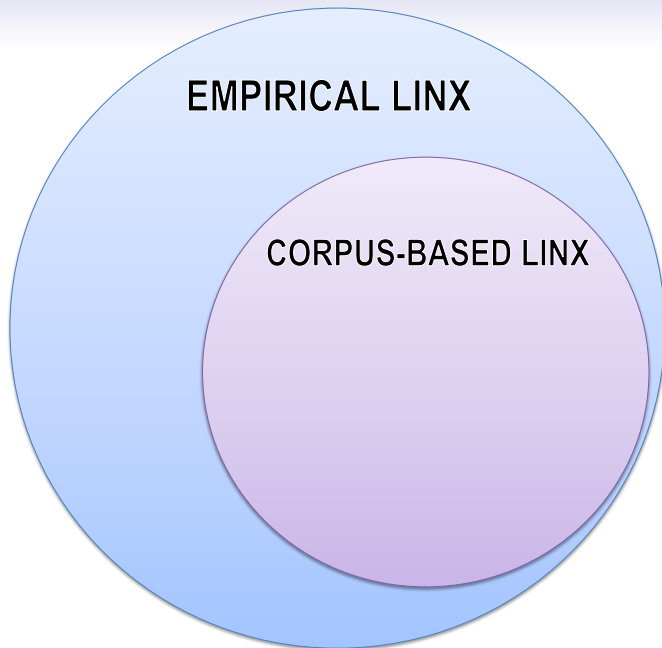
(Meyer 2002: 1)

the variationist method is a proper subset of the corpus-linguistic family of methods

“a corpus of language is a body of language occurring in a particular context”

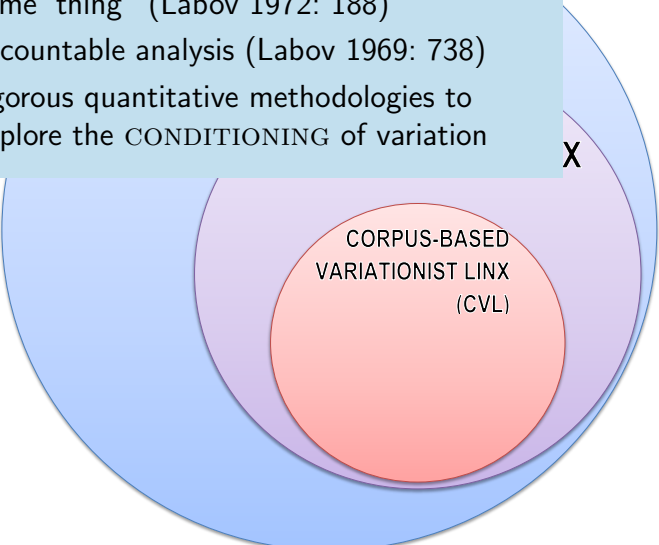
(McEnery et al. 2006: 4)

EMPIRICAL LINX

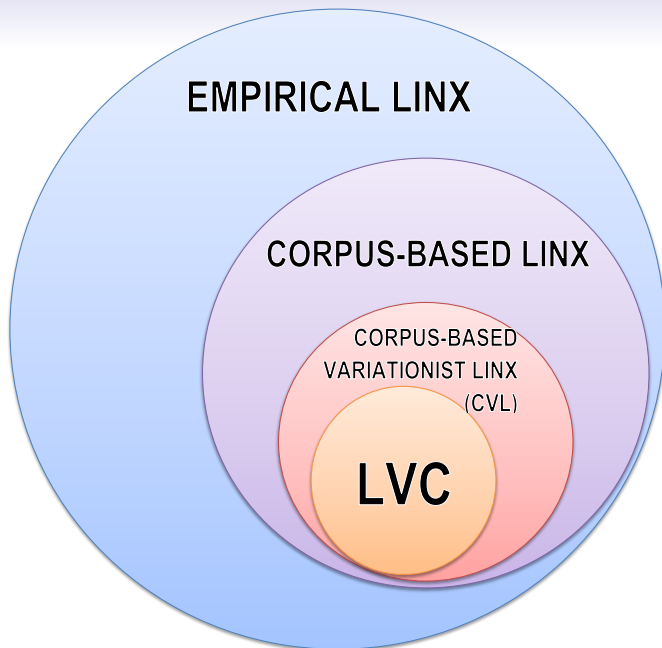


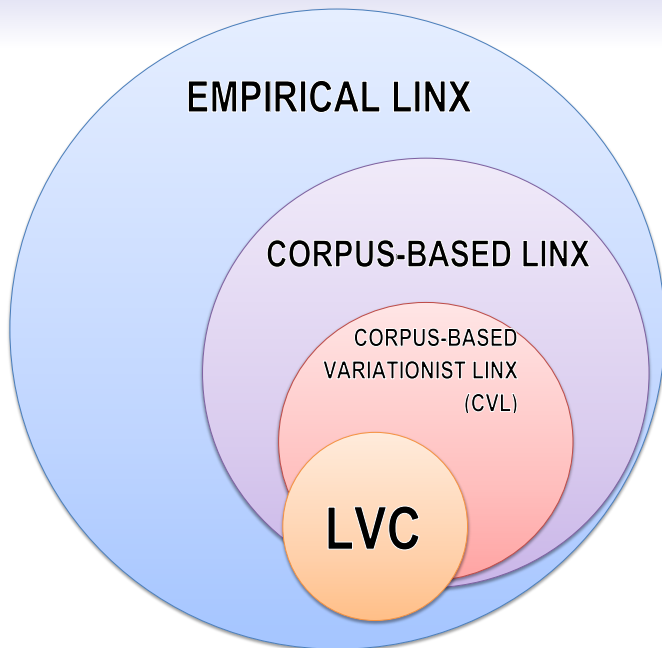
Defining CVL

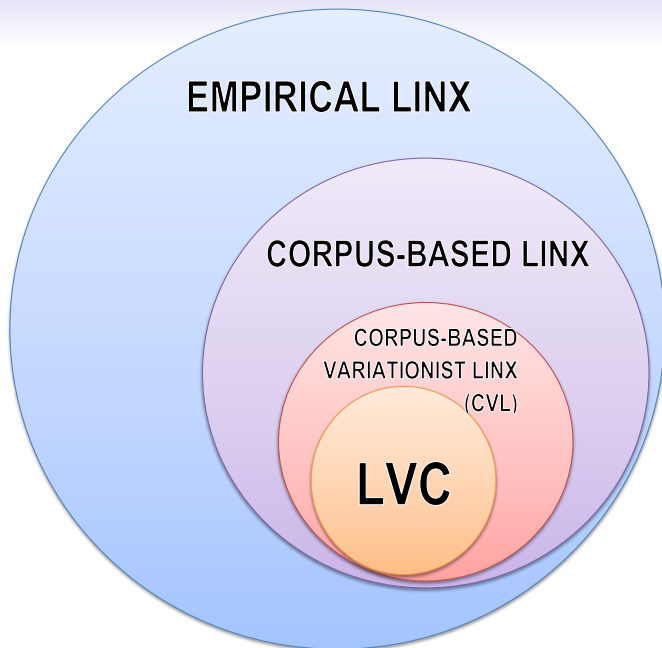
1. interest in “alternate ways of saying ‘the same’ thing” (Labov 1972: 188)
2. accountable analysis (Labov 1969: 738)
3. rigorous quantitative methodologies to explore the **CONDITIONING** of variation

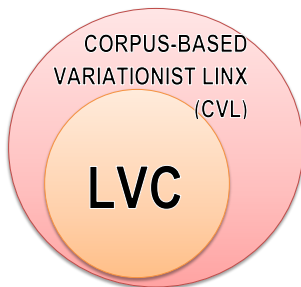


CORPUS-BASED
VARIATIONIST LINX
(CVL)









CVL: Who's out

- empirical but not corpus-based
(e.g. experimental psycholinguistics – Bock 1986)
- corpus-based/corpus-driven but not concerned with variation
(e.g. Rayson, Piao, Sharoff, Evert, and Moirón 2010, “Multiword expressions: hard going or plain sailing?”)
- corpus-based & concerned with variation but not using the variationist method
(e.g. Biber 1988)

CVL studies that fit the bill

Bresnan, Cueni, Nikitina, and Baayen (2007); Claes (2014); De Cuypere and Verbeke (2013); Ehret, Wolk, and Szmrecsanyi (2014); Grafmiller (2014); Gries (2005); Grondelaers and Speelman (2007); Heylen (2005); Hilpert (2008); Hinrichs and Szmrecsanyi (2007); Jaeger (2006); Levshina, Geeraerts, and Speelman (2013); Lohmann (2011); Pijpops and Van de Velde (2014); Schilk, Mukherjee, Nam, and Mukherjee (2013); Shih, Grafmiller, Futrell, and Bresnan (2015); Theijssen, ten Bosch, Boves, Cranen, and van Halteren (2013); Wolk, Bresnan, Rosenbach, and Szmrecsanyi (2013); Wulff, Lester, and Martinez-Garcia (2014), ...

Six differences between LVC and CVL



1. Focus on demographic factors

- **LVC**: focus on demographic factors (age, gender, . . .)
- **CVL**: more interested in macrosociological drifts/phenomena
(colloquialization, prescriptivism, standardization. . .)

2. Focus on phonetic variation

- **LVC**: dominated by work on phonetic variation
but see e.g. Weiner and Labov (1983); Tagliamonte et al. (2005);
Poplack and Dion (2009) ...
- **CVL**: tends to prioritize morphological, syntactic, or
lexical variation
but see e.g. Rosenfelder (2009)

3. Focus on vernacular speech

- **LVC**: especially interested in vernacular speech as manifested in sociolinguistic interviews (often enriched by data on style-shifting)
see Chambers (2003: 6)
- **CVL**: considerably less selective – in fact, many standard corpora sample multiple genres
(for example, the International Corpus of English covers 32 text types: e.g. face-to-face conversations, legal cross-examinations, business letters ...)

4. Focus on changes in apparent time

- **LVC**: apparent-time construct very popular
see Bailey et al. (1991)
- **CVL**: focus on changes in real time, drawing on increasingly massive historical corpora typically sampling a variety of written text types
see e.g. Hackert (next session), Raumolin-Brunberg (2005)

5. Theoretical orientation

Most CVL practitioners will identify as **usage-based linguists** in the following sense:

grammar is the **cognitive organization of one's experience with language** [...] certain facets of linguistic experience, such as the frequency of use of particular instances of constructions, have an impact on representation [...]

(Bybee 2006: 711; emphasis mine)

6. Cultural differences

- **fieldwork** – big role in LVC
- **coding and annotation** – LVC analysts not afraid of meticulous manual data analysis; CVL analysts more enthusiastic about using (semi-)automatic retrieval and annotation procedures
- **terminology**: “conditioning factor” vs “predictor”, “variant rate” vs “relative frequency”, etc.
- in the LVC community, keen awareness of and insistence on **foundational principles**

Cross-pollination potential



Fields of interest

1. Multi-variable studies
2. Research on register-induced variation
3. Probabilistic Grammar studies

Multi-variable studies



One variable at a time?

- one-variable-at-a-time methodology customary in LVC
(but see e.g. Corrigan et al. 2014)
- but recent interest in the joint behavior of multiple variables
(see Guy 2013)
- feature aggregation has been a theme in the corpus-linguistic literature for a long while
(Biber 1988)

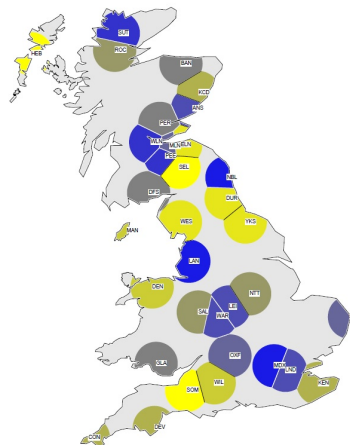
Szmrecsanyi (2013)

- “Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry”
- analyzes transcribed interviews sampled in the *Freiburg Corpus of English Dialects* to uncover big-picture geolinguistic patterns
(www.helsinki.fi/varieng/CoRD/corpora/FRED/)
- **dialectometry**: joint frequency variation of 57 morphosyntax features in 34 British English dialects

Regionally distinctive feature bundles – PC2

PC 2: Rotated component loadings.

[13]	the primary verb TO DO	.80
[15]	the primary verb TO HAVE	.80
[6]	<i>them</i>	.68
[25]	marking of epistemic and deontic modality: HAVE TO	.58
[34]	negative contraction	.58
[53]	zero complementation after THINK, SAY, and KNOW	.56
[39]	non-standard verbal -s	-.18
[44]	non-standard WAS	-.32



Map C.26 Principal Component 2. Mean component score map. Yellowish hues indicate higher mean component scores; bluish hues indicate lower mean component scores.

Hinrichs, Szmrecsanyi, and Bohmann (in press)

- (1)
- a. Tom saw the car **that** Mary had sold
 - b. Tom saw the car **which** Mary had sold
 - c. Tom saw the car Mary had sold

in written English, this variation is
undergoing massive shift from *which*
to *that*, spearheaded by AmE

Two candidate explanations

1. **prescriptivism**: “Careful writers [...] go *which*-hunting, remove the defining *whiches*, and by so doing improve their work”
(see Strunk and White 1999: 59)
2. the **colloquialization** of the norms of written English (Mair 2006: 88): *that* is the informal & vernacular variant (e.g. Tagliamonte et al. 2005)

Study design

- study \approx 17k RRCs and annotate for language-internal & and language-external predictors, as well as for **additional variables regulated by prescriptivism** as IVs:
 1. usage of passive voice
 2. preposition stranding
 3. split infinitives
 4. *shall* versus *will*
- regression to check extent to which the above features predict choice of relativizer
 - ⇒ hypothesis: if *that*-shift is prescriptivism-fueled, *which*-hunters should also comply with other precepts
- *that*-shift: **institutionally backed colloquialization**

The forests behind the trees

- single-variable studies fine if focus is really on the variables/variants (“trees”)
- but inadequate if is multidimensional lects (the “forests”) or drifts (colloquialization, . . .) which are of interest (see Nerbonne 2009 for discussion)
- aggregational methods fairly well-developed in the corpus-based literature

Research on register-induced variation

Register variation

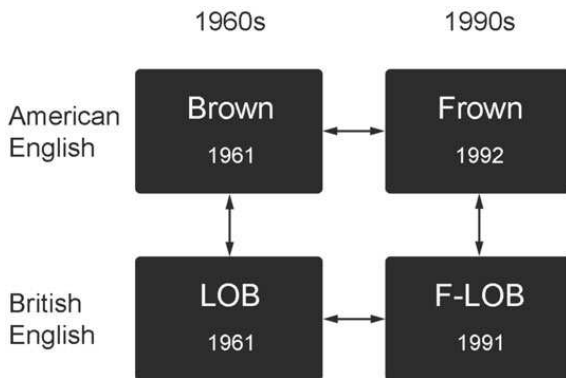
- vernacular speech as the register/style where variation is at its most interesting?
(see D'Arcy and Tagliamonte 2015 for critical discussion)
- long-standing corpus-linguistic interest in register differences
(consider work by Douglas Biber and collaborators)
- but the difference that register makes still under-researched in an explicitly variationist perspective

Ruette, Ehret, and Szmrecsanyi (to appear)

- how is lexical variation in standard English patterned in space, time, and across registers?
- draw on Semantic Vector Space modeling to create an unbiased lexical variable set ($N = 303$)
(e.g. *holiday-trip*, *sea-ocean*, *computer-pc*, ...)
- use aggregational techniques to rank lectal dimensions in terms of how strongly they trigger variation

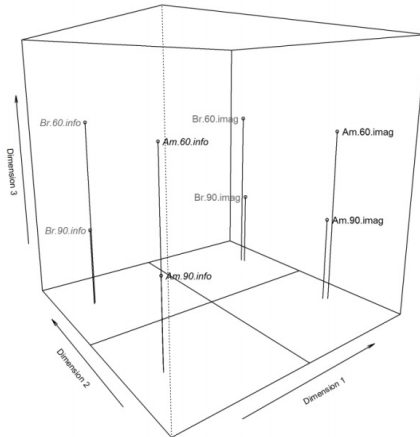
Data source: the Brown-family of corpora

Four corpora with (near-)identical design sampling written Standard English (1 million words each):



(see Hinrichs et al. 2010)

Individual Differences Scaling



Ranking of lectal dimensions

1. **register**
(info vs imaginative)
2. **variety**
(Br vs Am English)
3. **real time**
(1960s vs 1990s)

Figure 2: Group Stimulus Space showing the position of the subcorpora in three dimensions.

Grafmiller (2014)

- about the extent to which the probabilistic grammar of genitive choice differs across genres/registers
- (2)
 - a. [the Grizzlies]' [winning streak]
(the *s*-genitive)
 - b. [the sidekick] of [Gene Autry]
(the *of*-genitive)
- 9 predictors, 6 registers/genres (conversation, learned writing, non-fiction, general fiction, western fiction, press)
 - corpora: Switchboard/Brown

Language-internal predictors considered (model 1)

possessor animacy
rhythm
final sibilancy
possessor givenness
semantic relation
possessor/possessum length
type-token ratio
possessor text frequency
preceding genitive

Language-internal predictors considered (model 1)

Lots of interactions

the probabilistic grammar of genitive
choice is massively sensitive to genre
effects!

type-token ratio
possessor text frequency
preceding genitive

The importance of considering register

- corpus research: register is an extremely important language-external determinant of variation
- the plasticity of linguistic choice-making as a function of register remains comparatively under-researched
- new applications for the comparative sociolinguistics method?

Probabilistic Grammar studies



Preliminaries

- focus on variation-centered work
(e.g. Bresnan 2007; Bresnan and Ford 2010)
 1. syntactic variation – and change – is **subtle, gradient & probabilistic** rather than categorical in nature
(Bresnan and Hay 2008)
 2. linguistic knowledge includes **knowledge of probabilities**, and speakers have powerful predictive capacities
(see also Gahl and Garnsey 2004; Gahl and Yu 2006)

Methodology

adopt the variationist methodology and restrict attention to semantically equivalent ways of saying the same “thing”

(Labov 1972: 188)

(3) the dative alternation in English

- a. We sent [the president]_{recipient} [a letter]_{theme}
(the ditransitive dative)
- b. We sent [a letter]_{theme} to [the president]_{recipient}
(the prepositional dative)

Bresnan, Cueni, Nikitina, and Baayen (2007), based on meticulous annotation & regression analysis: ≈ 10 constraints

A dative model (based on Switchboard corpus data)

Probability of the prepositional dative = $1 / 1 + e^{-(X\beta + u_i)}$

where

$$\begin{aligned} \hat{X\beta} = & 1.1583 \\ & -3.3718 \{\text{pronominality of recipient} = \text{pronoun}\} \\ & +4.2391 \{\text{pronominality of theme} = \text{pronoun}\} \\ & +0.5412 \{\text{definiteness of recipient} = \text{indefinite}\} \\ & -1.5075 \{\text{definiteness of theme} = \text{indefinite}\} \\ & +1.7397 \{\text{animacy of recipient} = \text{inanimate}\} \\ & +0.4592 \{\text{number of theme} = \text{plural}\} \\ & +0.5516 \{\text{previous} = \text{prepositional}\} \\ & -0.2237 \{\text{previous} = \text{none}\} \\ & +1.1819 \cdot [\log(\text{length}(\text{recipient})) - \log(\text{length}(\text{theme}))] \\ \text{and } \hat{u}_i \sim & N(0, 2.5246) \end{aligned}$$

Figure 1. The model formula for datives

(Ford and Bresnan 2013)



The 100-split task

participants rate the naturalness of alternative forms as continuations of a context by distributing 100 points between the alternatives. Thus, for example, participants might give pairs of values to the alternatives like 25–75, 0–100, or 36–64. From such values, one can determine whether the participants give responses in line with the probabilities given by the model and whether people are influenced by the predictors in the same manner as the model.

(Ford and Bresnan 2013)



The 100-split task: an example

I'm in college, and I'm only twenty-one but I had a speech class last semester, and there was a girl in my class who did a speech on home care of the elderly. And I was so surprised to hear how many people, you know, the older people, are like, fastened to their beds so they can't get out just because, you know, they want
they get the wrong medicine, just because
aides or whatever

(1) just give them the wrong medicine

(2) give the wrong medicine to them

Predictions

the model suggests a 98–2 split in favor of the ditransitive dative in (1) – speakers tend to agree!

Some interesting Probabilistic Grammar work

- [Bresnan and Hay \(2008\)](#):
US-NZ differences
- [de Marneffe, Grimm, Arnon, Kirby, and Bresnan \(2012\)](#):
development of probabilistic grammars in children
- [Wolk, Bresnan, Rosenbach, and Szmrecsanyi \(2013\)](#):
real-time dynamics of probabilistic change
- [Grafmiller \(2014\)](#):
register-induced variation
- [Szmrecsanyi, Grafmiller, Heller, and Röthlisberger \(t.a.\)](#):
scope & limits of syntactic variation in varieties of English around the world



Around the world in three alternations

- project “Exploring probabilistic grammar(s) in varieties of English around the world”
(see <http://tinyurl.com/ng8ws6o>)
- main goal: understand the plasticity of probabilistic knowledge of English grammar, on the part of language users with diverse regional and cultural backgrounds

The particle placement alternation

- (4)
- a. The president looked_{verb} [the word]_{NP} up_{particle}
(V-DO-P – split pattern)
 - b. The president looked_{verb} up_{particle} [the word]_{NP}
(V-P-DO – unsplit pattern)

Particle placement: length effects are variable

(look up [the difficult word] vs look [the difficult word] up)

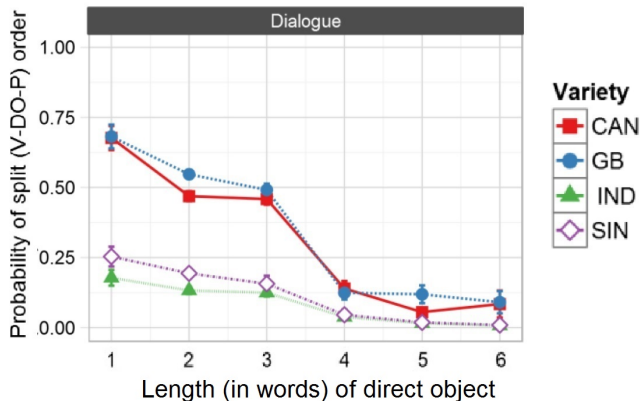


Figure: Predicted probabilities obtained from Conditional Random Forest model on corpus data (with 95% confidence intervals)

Why interesting?

- key interest in what language users know about the effect of language-internal constraints on grammatical variation (often as a function of language-external factors)
- methodological compatibility
- “balanced diet” (Guy 2014: 59) consisting of (abstract) constraints plus usage & experience

Concluding remarks



Conclusion

- corpus-based variationist linguistics (CVL) is compatible with LVC ...
- ... to the extent that we do not insist that variationist work must necessarily consider demographic factors such as age, gender, etc.
- cross-pollination potential

Thank you!

`benszm@kuleuven.be`

`http://www.benszm.net/`



References I

- Bailey, G., T. Wikle, J. Tillery, and L. Sand (1991, October). The apparent time construct. *Language Variation and Change* 3(03), 241.
- Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology* 18, 355–387.
- Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston and W. Sternefeld (Eds.), *Roots: Linguistics in Search of Its Evidential Base*, pp. 75–96. Berlin: Mouton de Gruyter.
- Bresnan, J., A. Cueni, T. Nikitina, and H. Baayen (2007). Predicting the Dative Alternation. In G. Boume, I. Kraemer, and J. Zwarts (Eds.), *Cognitive Foundations of Interpretation*, pp. 69–94. Amsterdam: Royal Netherlands Academy of Science.
- Bresnan, J. and M. Ford (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1), 168–213.
- Bresnan, J. and J. Hay (2008, February). Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua* 118(2), 245–259.
- Bybee, J. L. (2006). From Usage to Grammar: The Mind's Response to Repetition. *Language* 82(4), 711–733.

References II

- Chambers, J. K. (2003). *Sociolinguistic theory: linguistic variation and its social significance* (2nd ed ed.). Number 22 in *Language in society*. Oxford ; Malden, MA: Blackwell.
- Claes, J. (2014, July). A Cognitive Construction Grammar approach to the pluralization of presentational haber in Puerto Rican Spanish. *Language Variation and Change* 26(02), 219–246.
- Corrigan, K. P., A. Mearns, and H. Moisl (2014, January). Feature-based versus aggregate analyses of the DECTE corpus: Phonological and morphological variability in Tyneside English. In B. Szmrecsanyi and B. Wälchli (Eds.), *Aggregating Dialectology, Typology, and Register Analysis*. Berlin, Boston: DE GRUYTER.
- D'Arcy, A. and S. A. Tagliamonte (2015, October). Not always variable: Probing the vernacular grammar. *Language Variation and Change* 27(03), 255–285.
- De Cuypere, L. and S. Verbeke (2013, June). Dative alternation in Indian English: A corpus-based analysis. *World Englishes* 32(2), 169–184.
- de Marneffe, M.-C., S. Grimm, I. Arnon, S. Kirby, and J. Bresnan (2012, January). A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes* 27(1), 25–61.
- Ehret, K., C. Wolk, and B. Szmrecsanyi (2014). Quirky quadratures: on rhythm and weight as constraints on genitive variation in an unconventional data set. *English Language and Linguistics* 18(02), 263–303.

References III

- Ford, M. and J. Bresnan (2013). Studying syntactic variation using convergent evidence from psycholinguistics and usage. In M. Krug and J. Schlüter (Eds.), *Research Methods in Language Variation and Change*. Cambridge: Cambridge University Press.
- Gahl, S. and S. Garnsey (2004). Knowledge of Grammar, Knowledge of Usage: Syntactic Probabilities Affect Pronunciation Variation. *Language* 80, 748–775.
- Gahl, S. and A. C. Yu (2006). *Special theme issue: Exemplar-based models in linguistics*. The linguistic review. Mouton de Gruyter.
- Grafmiller, J. (2014, November). Variation in English genitives across modality and genres. *English Language and Linguistics* 18(03), 471–496.
- Gries, S. T. (2005). Syntactic Priming: A Corpus-based Approach. *Journal of Psycholinguistic Research* 34(4), 365–399.
- Grieve, J. (2011). A regional analysis of contraction rate in written Standard American English. *International Journal of Corpus Linguistics* 16(4), 514–546.
- Grondelaers, S. and D. Speelman (2007, January). A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory* 3(2).
- Guy, G. R. (2013, June). The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables? *Journal of Pragmatics* 52, 63–71.

References IV

- Guy, G. R. (2014, April). Linking usage and grammar: Generative phonology, exemplar theory, and variable rules. *Lingua* 142, 57–65.
- Heylen, K. (2005). A Quantitative Corpus Study of German Word Order Variation. In S. Kepser and M. Reis (Eds.), *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, pp. 241–264. Berlin, New York: Mouton de Gruyter.
- Hilpert, M. (2008, November). The English comparative – language structure and language use. *English Language and Linguistics* 12(03), 395.
- Hinrichs, L., N. Smith, and B. Waibel (2010). Manual of information for the part-of-speech-tagged, post-edited "Brown" corpora. *ICAME Journal* 34, 189–231.
- Hinrichs, L. and B. Szmrecsanyi (2007, November). Recent changes in the function and frequency of Standard English genitive constructions: a multivariate analysis of tagged corpora. *English Language and Linguistics* 11(03), 437–474.
- Hinrichs, L., B. Szmrecsanyi, and A. Bohmann. Which-hunting and the Standard English Relative Clause. *Language* 91(4).
- Jaeger, T. F. (2006). *Redundancy and Syntactic Reduction in Spontaneous Speech*. PhD Thesis, Stanford University.
- Kennedy, G. (1998). *An introduction to corpus linguistics*. Studies in language and linguistics. London: Longman.
- Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language* 45, 715–762.



References V

- Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Philadelphia press.
- Levshina, N., D. Geeraerts, and D. Speelman (2013, June). Towards a 3d-grammar: Interaction of linguistic and extralinguistic factors in the use of Dutch causative constructions. *Journal of Pragmatics* 52, 34–48.
- Lohmann, A. (2011, October). Help vs help to: a multifactorial, mixed-effects account of infinitive marker omission. *English Language and Linguistics* 15(03), 499–521.
- Mair, C. (2006). *Twentieth-century English: History, variation, and standardization*. Cambridge: CUP.
- McEnery, T., R. Xiao, and Y. Tono (2006). *Corpus-based language studies: an advanced resource book*. New York: Routledge.
- Meyer, C. F. (2002). *English corpus linguistics: an introduction*. Studies in English language. Cambridge, UK ; New York: Cambridge University Press.
- Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics Compass* 3(1), 175–198.
- Pijpops, D. and F. Van de Velde (2014, January). A multivariate analysis of the partitive genitive in Dutch. Bringing quantitative data into a theoretical discussion. *Corpus Linguistics and Linguistic Theory* 0(0).
- Poplack, S. and N. Dion (2009). Prescription vs. praxis: The evolution of future temporal reference in French. *Language* 85(3), 557–587.

References VI

- Raumolin-Brunberg, H. (2005, March). The diffusion of subject YOU: A case study in historical sociolinguistics. *Language Variation and Change* 17(01).
- Rayson, P., S. Piao, S. Sharoff, S. Evert, and B. V. Moirón (2010, April). Multiword expressions: hard going or plain sailing? *Language Resources and Evaluation* 44(1-2), 1–5.
- Rosenfelder, I. (2009). *Sociophonetic variation in educated Jamaican English: An analysis of the spoken component of ICE-Jamaica*. PhD dissertation, University of Freiburg, Freiburg.
- Ruette, T., K. Ehret, and B. Szmrecsanyi. A lectometric analysis of aggregated lexical variation in written Standard English with Semantic Vector Space models. *International Journal of Corpus Linguistics*.
- Schilk, M., J. Mukherjee, C. Nam, and S. Mukherjee (2013, January). Complementation of ditransitive verbs in South Asian Englishes: a multifactorial analysis. *Corpus Linguistics and Linguistic Theory* 9(2).
- Shih, S., J. Grafmiller, R. Futrell, and J. Bresnan (2015, January). Rhythm's role in genitive construction choice in spoken English. In R. Vogel and R. Vijver (Eds.), *Rhythm in Cognition and Grammar*. Berlin, München, Boston: DE GRUYTER.
- Strunk, W. and E. B. White (1999, September). *The Elements of Style* (4th ed.). Longman.
- Szmrecsanyi, B. (2013). *Grammatical variation in British English dialects: a study in corpus-based dialectometry*. Cambridge, New York: Cambridge University Press.

References VII

- Szmrecsanyi, B., J. Grafmiller, B. Heller, and M. Röthlisberger. Around the world in three alternations: modeling syntactic variation in varieties of English. *English World-Wide* 37(2).
- Tagliamonte, S., J. Smith, and H. Lawrence (2005). No taming the vernacular! Insights from the relatives in northern Britain. *Language Variation and Change* 17(1), 75–112.
- Theijssen, D., L. ten Bosch, L. Boves, B. Cranen, and H. van Halteren (2013, January). Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory* 9(2), 227–262.
- Weiner, J. and W. Labov (1983). Constraints on the agentless passive. *Journal of Linguistics* 19, 29–58.
- Wolk, C., J. Bresnan, A. Rosenbach, and B. Szmrecsanyi (2013, January). Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3), 382–419.
- Wulff, S., N. Lester, and M. T. Martinez-Garcia (2014, June). That-variation in German and Spanish L2 English. *Language and Cognition* 6(02), 271–299.